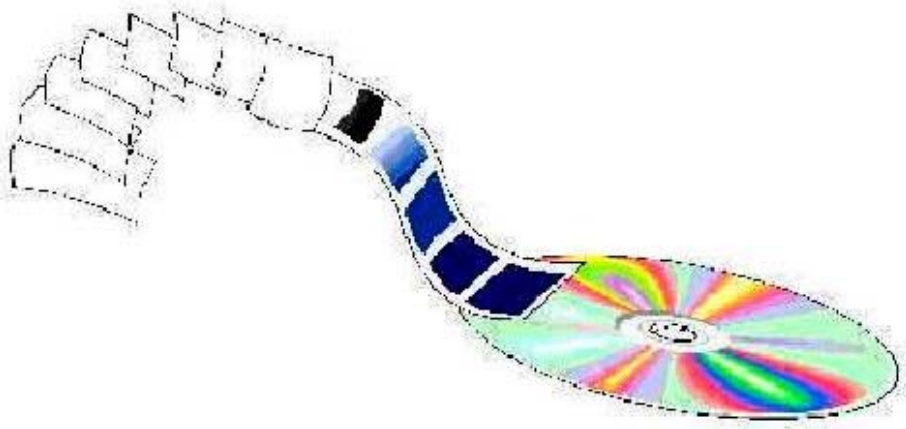
Document Imaging Technology for Arizona Government Records



Arizona State Library, Archives and Public Records
Records Management Division

July 2002

State Records Management Center
1919 West Jefferson Street
Phoenix, Arizona 85009
Phone: 602-542-3741
Fax: 602-542-3890
website: www.lib.az.us
e-mail: rmd@lib.az.us

PROVIDING ACCESS

Preserving Arizona

Chapter 1 – Film-based Imaging

A BRIEF HISTORY

Film-based document imaging is an old technology predating the practical use of electricity and, in fact, predating the office typewriter. In 1853 an Englishman named John Dancer photographically reduced a 20 inch document to 1/8 inch. He used a 100x microscope to read his new micro-images.

Film-based imaging remained a novelty until the 1860's when a Frenchman, Rene Dagron, patented a process. Then, during the siege of Paris in the Franco-Prussian War, (c.1870) microphotography was used to reduce important documents to a size small enough where several could be placed in a capsule on a pigeon's leg. These capsules were the basic means of communications in and out of Paris during the siege.

In 1925 an American banker, George L. McCarthy, patented the first Rotary Camera which soon went into use filming checks in New York City. The first readers came into use in 1928 making microfilm more easily usable and convenient.

John Langan, an American working for the O.S.S. (Army Office of Strategic Services), developed the aperture card in 1943. The aperture card has since become the standard microform for engineering drawings and is one of the most widely used forms of film-based imaging. It is also the last use of the punch-card, once the primary input for computers.

Computer output microfilm (COM) dates back to the 1950's. COM was originally developed by Stromborg Carlson Corporation for the U.S. Department of Defense. It has since become one of the most popular and cost effective forms of film-based imaging.

The 1970's and 1980's were the glory years for film-based imaging. Computer assisted retrieval (CAR) roll film systems were very popular for large volumes of active records. Computer output microfilm (COM), usually in fiche form, was extremely popular as an alternative to printed paper computer generated reports.

The film-based imaging industry underwent major changes in the 1990's. With increasing competition from digital document imaging and the decreasing cost of computer memory, the number of manufacturers and service providers substantially decreased. Likewise, the number of products available also decreased.

Is microfilm still a viable records medium? The answer is a qualified yes. Microfilm remains an excellent preservation medium for long term and valuable records. It is a relatively inexpensive reproduction medium for duplication of records required for a disaster prevention program. It is also very appropriate for collections of records and publications that have historical value, since the original silver halide master film can be stored for hundreds of years without appreciable degradation.

DEFINITION OF FILM-BASED IMAGING

Film-based imaging is the technique(s) associated with the production, handling and use of various forms, usually made of photographic film, which contain images of information too small to be read without magnification.

Film-based imaging is a broader term but is frequently synonymous with "micrographics," "microfilming," "microphotography," and "micro-reproduction." Electronic document imaging is covered in Chapter 2.

MICROFORMS

A microform is the form of photographic media containing micro-images. Of the many forms once available only roll film and fiche survived into the 21st century. Many microforms have been available over the years with seven forms being most commonly used: (See Figure 1):

- **Roll Film** - Also called reels, this is one of the oldest forms and is most commonly available in 16mm and 35mm widths. 16mm microfilm is frequently used for security or archival filming of smaller documents including checks, office files and correspondence. 35mm microfilm is usually used when large documents or publications are being filmed. Newspapers, maps and security film of engineering drawings are common 35mm applications.
- **Cassette** - This is a double core self-contained unit for 16mm roll microfilm. It is somewhat similar in construction to a magnetic tape cassette but considerably larger (6 in. x 5 in. 1 in.). A cassette does not require threading onto a reader and need not be rewound after use. This form lost popularity in the 1980's and was virtually out of use by the 1990's.
- **Cartridge** - This had been the most popular small document 16mm microform in use. Several types of proprietary single core enclosed units were available for 16mm roll film. They were all approximately 4 in. x 4 in. x 1 in. in size and made of hard plastic. Threading and drive mechanisms varied from type to type, but some readers are adaptable to more than one type of cartridge. The ANSI (American National Standards Institute) standard cartridge currently accounts for most cartridges still in use.
- **Aperture Card** - Essentially a Hollirith Card (i.e. IBM-card; tab-card, punch-card) containing an aperture or window into which a single frame of 35mm microfilm is inserted. This is the most common microform for engineering drawings.
- **Tab-Jac** - This is a paper card usually either Hollirith Card size or 148mm x 105mm (approx. 6 in. X 4 in.) into which channels have been cut. Thin polyester film is applied to either side of the channels enabling

strips of 16mm or 35mm microfilm to be inserted. The paper may be preprinted and may have eye readable information written onto the card.

- **Jacket** - Polyester film laminated to form a clear plastic card with channels, into which strips of 16mm or 35mm microfilm may be inserted. Jackets are usually 148mm x 105mm with a frosted strip at the top to accept typed or written labeling.
- **Fiche** - A sheet of photographic film containing micro-images in a grid format. Most commonly 148mm x 105mm, fiche may contain either source document images or computer generated characters or images (COM).

THE NORSAM PANCAKE DISC™

A relatively new technology, originally licensed from and developed by scientists at Los Alamos National Laboratory, is currently being marketed by Norsam Technologies, Inc. This product is in very limited use, but holds some real promise for very high density storage of analog images for very long periods of time (advertised for 1000+ years).

This new technology uses a focused ion beam machine to mill the images into the surface of a metallic nickel "HD-Rosetta Disc." This machine essentially acts as a printer printing digitally scanned bimodal text images at 300 dpi (compared to original source). Graphic material is printed as a dithered grayscale at 100 dpi (compared to original source). The major innovation is that this *printer* is capable of printing an 8.5 inch by 11 inch document as small as 200 by 300 microns (approximately a 600X reduction).

The disc is then stored in a sealed "encapsulator box" containing an inert gas such as nitrogen. This process is currently intended for high value documents requiring very long preservation periods.

SOURCE DOCUMENT MICROFILMING

When original paper documents are recorded onto microfilm the process and product are referred to as "source document." This is a very general term which is used whenever paper documents are being filmed.

Source document filming may include any of the microforms previously described. The individual characteristics of the records to be microfilmed and their use will dictate the microform to be used. These characteristics include:

1. Document size
2. Filing sequence
3. Retrieval activity
4. Historical value
5. Retention requirements

COMPUTER OUTPUT MICROFILM (COM)

Microforms may be generated on-line directly from the computer or off-line from magnetic media (e.g.. tape, disc, etc.). The result is referred to as COM. *When COM is employed, paper does not serve as an intermediate document.*

The most common microform for COM is fiche. One 148mm x 105mm fiche may contain up to 644 pages of computer generated data. The images on the fiche may even have a printed form superimposed photographically so that the image looks identical to a computer run paper form.

Although not as common as fiche, COM has successful 16mm roll film and aperture card applications. Digitized engineering drawings and CAD (Computer Assisted Design) graphics are good COM aperture card applications.

A newer use for COM is to produce "archival quality" microfilm from the electronic/digital images stored in a document imaging system. This COM generally produces bi-tonal film images from tiff images stored in a computer system. The end product is 16 mm roll film with "blip" encoding. This microfilm, if properly processed and stored, meets the ANSI standards for long term microfilm.

COMPUTER ASSISTED RETRIEVAL (CAR)

Computers have been used as an indexing tool for film-based imaging since the mid-1960s. The first uses were crude, but they showed the great potential that existed for the merging of source document microfilm with an electronic database index.

Basically two types of CAR Systems were very popular in the 1980's and *may* even remain in use today:

1. Roll film (including cartridge and cassette) systems
2. Fiche Systems

Roll film CAR systems were by far the most common and popular type. They were particularly suitable to large files of document which were used as support for a computer file.

In general, CAR systems involved relatively expensive equipment and required considerable effort and quality control to ensure reliability. They afforded the user rapid retrieval from a large base of source document microfilm, and they increased overall productivity in large, very active file areas.

HYBRID IMAGING SYSTEMS

A recent development is called hybrid imaging systems that are made up of both micrographic and electronic/digital images. These systems are covered in the following chapter.

Chapter 2 - Electronic/Digital Imaging

BACKGROUND

As computers advanced, with more and more processing power, memory and high density data storage, digital imaging became a popular and very viable photographic option. In the broadest sense digital (or electronic) imaging includes all digital photographic techniques from medical imaging systems to digital color photography. Records managers are concerned primarily with *document imaging* which is the digital imaging of documents.

Document imaging, as all digital imaging, actually has its roots in the purely scientific arena where the techniques were used in the 1960s and 1970s to relay photographs to earth from various space probes. As computers became more powerful and storage media more efficient the technology spread into the business community and eventually included the imaging of documents.

In the mid-1980s various optical disk media became available paving the way for the document imaging systems we are currently familiar with. The early systems were all proprietary and expensive costing a minimum of about $250,000. These early systems and the software developed for them brought to us most of the features and advantages associated with document imaging.

In the 1990s the systems became more and more standardized and "open". The prices also continued to drop and the storage options changed from expensive proprietary optical disks to CD-R and high density magnetic storage.

THE 21ST CENTURY

As we begin the new century (millennium) document imaging has become relatively inexpensive and so standardized as to be primarily a factor of software applications. It is not the intent of this manual to delve into the technological aspects of imaging, but rather to discuss records applications. For more information on the technology of imaging, other sources, including publications from AIIM (Association for Information and Imaging Management) are suggested.

APPROPRIATE APPLICATIONS

Document imaging is a very attractive method for the storage, indexing and retrieval of document type records. It takes up virtually no space, in the traditional sense, and the images can be viewed on a standard computer display and printed on any modern laser or ink-jet printer.

The primary costs associated with document imaging are the scanning conversion costs and the indexing (data input) costs. These are labor intensive activities and can incur substantial expenses. It is important to analyze these costs into the total cost of an imaging solution.

Digital imaging may also enhance the "work-flow" of documents that are acted on through various steps in some procedure or system. Work-flow software typically routes the documents to their next step in the process after a particular step is completed.

INDEXING

Indexing the digitized documents in such a way as to find them when they are needed is one of the most important aspects of a document imaging system. Too frequently potential users get the impression that all they need to do is scan the documents and the computer will be able to automatically find the document they want it.

OCR (optical character recognition) is commonly available software that can actually identify letters and words when they are searched for. The OCR scan may take place at the same time the document is scanned for imaging or it may be a separate process. OCR works best on printed materials, but is still subject to errors. Hand written materials are very difficult to effectively scan with OCR. Most OCR retrieval software is based on key word or straight text retrieval. A retrieval based on the word "dog" will yield hits from "dog license" to "dog day afternoon." Some retrieval programs are capable of more advance searches by using multiple words or phrases. There is newer and more expensive retrieval software becoming available which is based on fuzzy logic, and is very effective for more advanced type searches.

More typically a database type search engine is used. These search engines are essentially databases including specific search fields that are associated with the imaged document. These search fields require data input and can be labor intensive, depending on how much *descriptive metadata* is entered.

Appropriate records applications are usually active records that need to be retrieved quickly and often. Various people may also need them at the same time

INAPPROPRIATE RECORDS

Some records are less appropriate for document imaging systems. These include:

- Case or project files that remain active over a prolonged period of time.
- Inactive records or files.
- Permanently retained records.
- Inactive records with very short retention periods.

Digital imaging should never be used as a storage medium for inactive records. Storing the paper will always be less costly than scanning and indexing the documents. And records destruction at the end of the retention period is much less complex for paper records.

HYBRID SYSTEMS

Like mules, navel oranges and patent roses, record-keeping systems can also be developed into hybrids. In general, hybrid imaging systems incorporate both microfilm and digital imaging for the purpose of realizing the advantages of both.

There are three methods to produce hybrid imaging systems:

1. Camera/scanners. Microfilm camera/scanners contain the essential elements of both a microfilm camera and a document scanner. In one pass a document is photographed onto silver halide microfilm and scanned to a TIFF image. The microfilm may be stored for a very long period of time, even permanently. The film provides a human readable long-term format, while the scanned image may be indexed and used in an automated image retrieval system.
2. Graphic COM. COM is computer Output Microfilm. Graphic COM is simply the output of images to microfilm. Specialized equipment is used to transfer the TIFF (or other format) images at a high rate of speed to the microfilm.
3. Microfilm scanners. Microfilm, microfiche and aperture cards may all be scanned using specialized equipment. The equipment is capable of performing a200 dpi to 400 dpi scan from the microfilm images. The speed of these scans is about one image per second.

Chapter 3 - Legality, Feasibility, Cost/Benefit Analysis

LEGALITY OF IMAGING

Generally speaking, film-based and electronic images are a legal form of documentation. Images and copies made from them are not excluded under the rules of evidence and are therefore admissible evidence in all courts, commissions or quasi-judicial bodies in Arizona, the U.S. Government and other state governments. However, the use of imaging may occasionally be limited by a specific statute or regulation.

ARIZONA REQUIREMENTS AND REQUIRED APPROVALS

Pursuant to A.R.S. §41-1348 records of state agencies or political subdivisions may be microfilmed or electronically imaged *only* after approval has been granted by the Arizona State Library, Archives and Public Records. To obtain approval a *Request for Document Imaging Implementation*, Figure 2, must be submitted to the State Library for each project being planned. This form is available in paper by calling the Records Management Division at 602-542-3741 or in PDF format at the State Library web site, www.lib.az.us.

In addition to this, and pursuant to A.R.S. §41-3504.(g), *executive branch state agencies* must receive approval from GITA (Government Information Technology Agency) for any technology projects with a total cost of $25,000 or more. For more information call GITA at 602-340-8538.

COMMON ASSUMPTIONS

There are many assumptions made about imaging which are just not true. They include:

- Imaging paper records is less expensive than storing them.
- Digital images are archival media and therefore will last forever.
- Microfilmed records must be retained permanently.
- The best reason for imaging is space savings.

"Conventional Wisdom" has taught us that the above statements are true. However, this *"Conventional Wisdom"* is not based on any real analysis of a situation. More often than not, these assumptions are not true.

NEW GENERALITIES

Based on analyses performed by Records Management Division (RMD) staff the following generalities may be considered valid:

- Records may be stored for up to 25 years in commercial record centers for cost of imaging them.
- Records may be stored in excess of 60 years in an in-house records center for the cost of imaging them.
- Microfilm will not last permanently unless **all** factors regarding archival quality film, processing and storage are met.
- The retention period for document images should not exceed that which was or would be established for equivalent paper records.

These new generalities are based upon actual practice, fact and analysis; and will remain true in most cases. However, certain requirements and situations may preclude them, and a specific study and analysis will be necessary.

FEASIBILITY OF IMAGING

Before embarking on a micrographics or imaging project, and indeed even before performing a cost/benefit analysis, it should be determined what type, if any, imaging system will perform well. Among the questions which must first be answered are:

- What are the physical characteristics of the documents to be filmed/imaged? (i.e. size, color, condition)
- Where and how are the records filed?
- What is the frequency of records retrieval?
- How are the records usually retrieved - by single specific document or by an entire case file?
- How are the records to be indexed for retrieval following imaging?
- What is the approved records retention period?
- What is the volume of records to be imaged?
- Will the records be updated following filming/imaging?
- Are any of the records to be filmed/imaged confidential?

When performing the feasibility study the following areas need to evaluated and analyzed (source: Georgia State Archives):

- Business Process Analysis
- Work Flow Evaluation
- A complete inventory of existing records
- Data needs assessment
- Network support
- Cost/benefits analysis
- Projected growth
- Retention and legal requirements

Once the feasibility of the imaging application is determined and the type or types of imaging to be considered is ascertained, the cost/benefit analysis should be undertaken.

COST/BENEFIT ANALYSIS

Simply put the cost/benefit analysis compares the cost of the present system to the cost of the proposed system. The analysis can easily be expanded to include more than one proposed system so as to provide for alternate choices.

Before any analysis can be made, however, a common basis for comparison of the systems must be defined. Volumes of recorded information, rates, of input and search time, and coverage of the present system should be compared against similar volume, rates, and coverage of proposed alternatives systems.

Adjustments to the data may be necessary to obtain a valid comparison if the present system and the proposed system do not appear to be basically similar in all these areas.

Comparing tangible costs is the easiest and most acceptable method of presenting such a cost/benefit analysis. By assigning dollar values to represent the costs of both systems and the benefits to be obtained from a proposed system, it is possible to derive comparable costs, net savings, or cost avoidance figures.

Direct labor costs are calculated on the basis of total hours required to operate the present information facilities and those estimated to operate a proposed system. The total hours are converted to dollars using appropriate average hourly wage rates paid to various types of employees involved.

(ERE) employee related expenses, (i.e. benefits, payroll taxes, retirement, FICA,etc.) are expressed as a percentage of the direct labor costs. This can usually be determined by contacting your agency fiscal officer.

Indirect labor includes supervisory and managerial costs, clerical support and other labor not directly related to performing the task being measured. Indirect labor is usually expressed as a percentage of direct labor costs.

Equipment costs include purchases (amortized), rentals, depreciation, and maintenance expressed in dollars and converted to annual, monthly, or hourly rates as appropriate to match other cost categories.

Consumables include supplies such as film, film processing, optical discs, reader-printer or laser printer paper and toner, film cartridges, projection lamps, labels, file folders, etc.

Overhead costs, such as space, management services, and heat and light is usually expressed as a percentage of direct labor costs. This overhead figure can represent part of the intangible costs of the present and proposed system.

It is more difficult to compute dollar benefits derivable from the installation of a new system. These are often intangible and speculative, such as faster retrieval of information, greater use of the data base, and better service to the public. There could even be some intangible costs of the present system that might not have been included in the overhead percentage; for example, estimated costs of "can't finds" or other costs associated with lack of retrieval of needed information because of system deficiencies. Such intangible benefits and costs can be estimated, quantified, and thoroughly documented.

In effect, this method provides a cost justification for a proposed system, which can be manipulated to show either a positive or negative relationship depending on what factors are considered. The systems designer must clearly establish in his presentation to agency managers the difference between the tangible and intangible costs and benefits. The more intangible the costs or benefits are, the less valid management should hold the dollar figures estimated to represent them. Often, the systems designer will attempt to convince agency managers that the intangible benefits obtained from a proposed microfilm/imaging system will outweigh the estimated greater tangible costs as compared to the present system.

VENDOR STUDIES

Vendor studies and proposals can be useful tools when evaluating a potential system. However, regardless of how "scientific" the study may appear its primary purpose is to sell. Never let a vendor study substitute for a good cost/benefit analysis.

Chapter 4 – Electronic/Digital Imaging Standards

(The standards in this document borrow significantly from "Electronic Document Imaging Systems Guidelines" published by the Georgia State Archives, Georgia Secretary of State.)

STANDARDS

The electronic/digital document imaging marketplace has been subject to rapid changes in technology, and the entry and exit of vendors. It is therefore necessary to have base standards to ensure the responsible implementation of these technologies. In order to secure approval from the State Library for a document imaging system an agency/office must meet the following standards:

"Permanent" records, pursuant to A.R.S. §39-101, must meet the *"Standards for Permanent Records Media and Storage"* developed and published by the Arizona State Library, Archives and Public Records. Electronic/digital document imaging systems do not currently meet these standards and should not be considered for preservation of permanent historical records.

An approved (pursuant to A.R.S.§§41-1347 and 41-1351) retention and disposition schedule is required for the records/documents being scanned into an electronic/digital document imaging system. A re-evaluation of retention requirements may be necessary when electronic/digital document imaging is being considered.

A practical effective migration/exit plan must be part of the system strategy for an electronic/digital document imaging system used for medium to long term records with a scheduled retention of 5 years or more. System administrators must plan to budget 5% to 10% of original system cost annually to cover the cost of upgrading and migration. Vendor software source code should be put into escrow to ensure the ability to migrate or exit from the system in the event of the vendor going out of business.

A records purging/destruction capability must be in place to ensure enforcement of approved records retention periods. Simply "deleting" image headers or pointers does not provide for thorough destruction of the images. Images of documents must be expunged or destroyed so as to leave no traces of them.

Hardware and software for the electronic/digital document imaging system must conform to nonproprietary standards and be constructed in open system architecture.

TIFF image format with standard nonproprietary image headers is preferred. "Open PDF" may also be considered. Under some circumstances other standard or cross-platform readable formats may be approved. Be aware that both GIF and most PDF formats, though widely used, are proprietary formats belonging to America Online, Inc. and Adobe Systems Inc. respectively. Images or headers may not be encrypted. Non-standard proprietary formats and headers must be avoided.

CCITT Group 3 or Group 4 shall be the standard compression for all TIFF images. Non-standard proprietary algorithms shall be avoided.

Thorough system documentation is necessary for :

- Hardware and software including brand names, version numbers, dates of installation, upgrades, etc.
- Data structure and content, including the file layout and data dictionaries
- Image enhancement algorithms.
- Operating procedures including scanning; data entry; revising, updating or expunging images; indexing; backups; and quality control.

Image resolution shall be between 200 and 400 dpi (dots per inch). Resolution below 200 dpi will not yield sufficient quality and resolution greater than 400 dpi is excessive for documentary materials.

Bimodal scans shall be used for typical documents. Grayscale scans should be avoided, because of the size of the resulting files, except where absolutely necessary for image quality.

Indexing must be appropriate for the records being imaged and based on realistic variables upon which the records will be requested. More advanced and specific metadata locators may be necessary for some types of records. Indexing must be accessible using standard SQL queries.

OCR (optical character recognition) may be used for indexing if appropriate and feasible for the types of records involved. OCR is subject to relatively high error rates and is dependent on fonts used and the quality of the print. "Full text" OCR scanning may also give a large number of false "hits" on searches for words because context is not usually considered. OCR also adds time to the scan process.

Labeling of image media is especially important when the images and the index are retained on separate media. Optical media used for image storage shall be labeled so as to identify the agency that produced the images, the records on the media and the system and software requirements to read the records.

Backup media shall be sufficiently labeled so as to identify the materials backed up. Backup media shall never be retained longer than the prescribed records retention period for the records on the backup. Backup media shall be appropriately stored with other computer system backups.

REQUEST FOR DOCUMENT IMAGING IMPLEMENTATION

Preserving Arizona's Access

Arizona State Library, Archives and Public Records
RECORDS MANAGEMENT DIVISION
1919 West Jefferson Street, Phoenix, Arizona 85009
Phone: 602-542-3741 ● Fax: 602-542-3890 ● E-mail: rmd@lib.az.us

Authorized pursuant to A.R.S. §41-1348
(Violation of this statute is a misdemeanor crime.)

New req. ❑ Revised ❑ Date Submitted

State Agency ❑ | Political Subdivision | Agency Name

Org. Unit/Division | Office | Phone

Address | City | AZ | Zip

Submitted By (Name) | Title | Signature X

DESCRIPTION OF RECORDS TO BE IMAGED: (Include record series name as it appears on the retention and disposition schedule and list various documents included in the series.)

Record Series	Retention (yrs.)

MICROFILM/FILM-BASED IMAGING (Briefly describe the filming application in the "Comments" area of this request.) ❑ | **ELECTRONIC/DIGITAL IMAGING** (Complete the balance of this request.) ❑

STUDIES PERFORMED: ❑ Feasibility (attach copy) ❑ Cost/Benefit (attach copy) ❑ P.I.J. (attach copy)

LIST HARDWARE BELOW:	LIST SOFTWARE BELOW:

MIGRATION/EXIT PLAN FOR LONG TERM RECORDS (Retention of 10 years or more)

❑ Migration/Exit Plan Adopted ❑ 5% – 10% System Cost annually Budgeted ❑ Vendor Source Code in Escrow

IMAGING SYSTEM HAS ABILITY TO COMPLETELY PURGE/DESTROY/EXPUNGE OBSOLETE RECORDS (IMAGES).	❑ Yes	❑ No
OPEN SYSTEM ARCHITECTURE	❑ Yes	❑ No
NONPROPRIETARY HARDWARE AND SOFTWARE	❑ Yes	❑ No

IMAGE FORMAT ❑ TIFF with Std. Headers ❑ Group 3 ❑ Group 4 ❑ Other:

IMAGE RESOLUTION ❑ 200 dpi ❑ 300 dpi ❑ 400 dpi ❑ Other: ❑ Bimodal ❑ Grayscale

INDEXING ❑ O.C.R. ❑ Manual ❑ Number of fields: **BACKUP MEDIA:**

COMMENTS:

Approved by: | Approval Date

Director, Arizona State Library, Archives and Public Records | Expires on (Approval date + 5 years)

RMC-2 R04/02

Figure 2